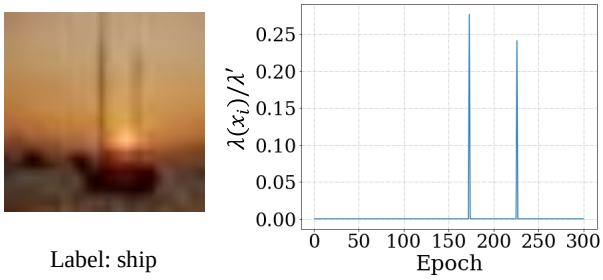
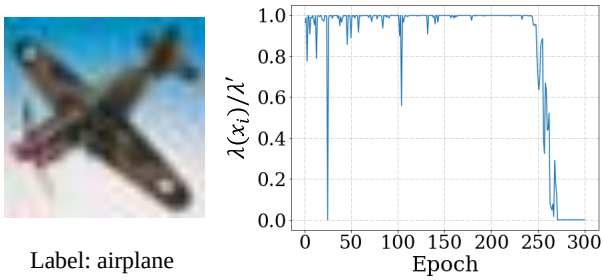
Manifold Regularized Dynamic Network Pruning (Supplementary Material)

Yehui Tang[1,2], Yunhe Wang[2]*, Yixing Xu[2], Yiping Deng[3], Chao Xu[1], Dacheng Tao[4], Chang Xu[4]
[1] Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University.
[2] Noah's Ark Lab, Huawei Technologies. [3] Central Software Institution, Huawei Technologies.
[4] School of Computer Science, Faculty of Engineering, University of Sydney.
yhtang@pku.edu.cn; yunhe.wang@huawei.com; c.xu@sydney.edu.au



(a) Complex Instance



(b) Simple instance

Figure S1. Coefficient $\lambda(\boldsymbol{x}_i)$ that controls network sparsity for complex and simple instances.

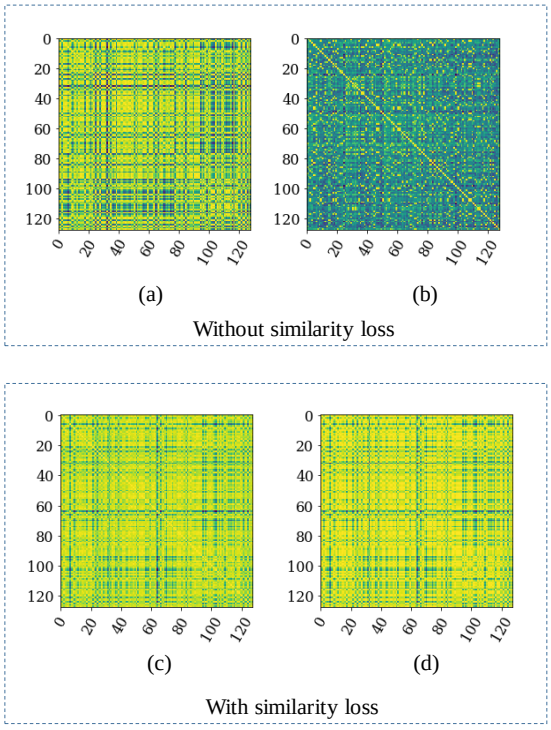


Figure S2. Similarity matrices. (a) Similarity matrix R^l for intermediate features without similarity loss. (b) Similarity matrix T^l for channel saliencies without similarity loss. (c) Similarity matrix R^l for intermediate features with similarity loss. (d) Similarity matrix T^l for channel saliencies with similarity loss.

1. Coefficient $\lambda(\boldsymbol{x}_i)$ for different instances

In the training procedure, the coefficient $\lambda(\boldsymbol{x}_i)$ (Eq. (6) of the main paper) controls the weight of sparsity loss according to the complexity of each instance. Recall that $\lambda(\boldsymbol{x}_i) = \lambda' \cdot \beta_i \frac{C - \mathcal{L}_{ce}(\boldsymbol{x}_i, \mathcal{W})}{C} \in [0, \lambda']$, where λ' is a fixed hyper-parameter for all instances. Using ResNet-56 as the backbone, the variable parts $\lambda(\boldsymbol{x}_i)/\lambda' \in [0, 1]$ for complex/simple examples in CIFAR-10 are shown in Figure S1. $\lambda(\boldsymbol{x}_i)/\lambda'$ keeps small for complex examples (*e.g.*, the vague 'ship' in (a)) and then less channels of the predefined networks are pruned for keeping their representation capabilities. When sending simple examples (*e.g.*, clear 'airplane' in (b)) to the dynamic network, $\lambda(\boldsymbol{x}_i)/\lambda'$ keeps large in most of the epochs, and thus the corresponding sub-network becomes sparser continuously as the numbers of iteration increases. Note that $\lambda(\boldsymbol{x}_i)/\lambda'$ changes dynamically in the training process. For example, the sparsity weight automatically decreases in the last few epochs as the corresponding sub-network is compact enough and should pay more attention to accuracy (Figure S1 (b)), which ensures that the models can fit input instances well.

*Corresponding author.

2. Similarity Matrices

The similarity matrices R^l for intermediate features and T^l for channel saliencies are shown in Figure S2, where different colors denote the degree of similarity (*i.e.*, a yellower point means higher degree of similarity between two instances). The ResNet-56 model trained with/without the similarity loss $\mathcal{L}_{sim}$ (Eq. (10) in the main paper) is used to generate features and channel saliencies for calculating the similarity between instances randomly sampled from CIFAR-10. When training dynamic network without similarity loss $\mathcal{L}_{sim}$, the similarity calculated by features and that by channel saliencies are very different (Figure S2 (a), (b)). When using similarity loss (Figure S2 (c), (d)), the similarity matrices R^l and T^l are more analogous as the similarity loss penalizes the inconsistency between similarity matrices to align the similarity relationship in the two spaces.

3. Visualization of instances with different complexity

We sample representative images with different complexity from ImageNet and intuitively show them in Figure S3. From top to bottom, the computational costs of sub-networks used to predict labels continue to increase. Intuitively, simple instances that can be accurately predicted by compacted networks usually contain clear targets, while the semantic information in complex images are vague and thus requires larger networks with powerful representation capability.

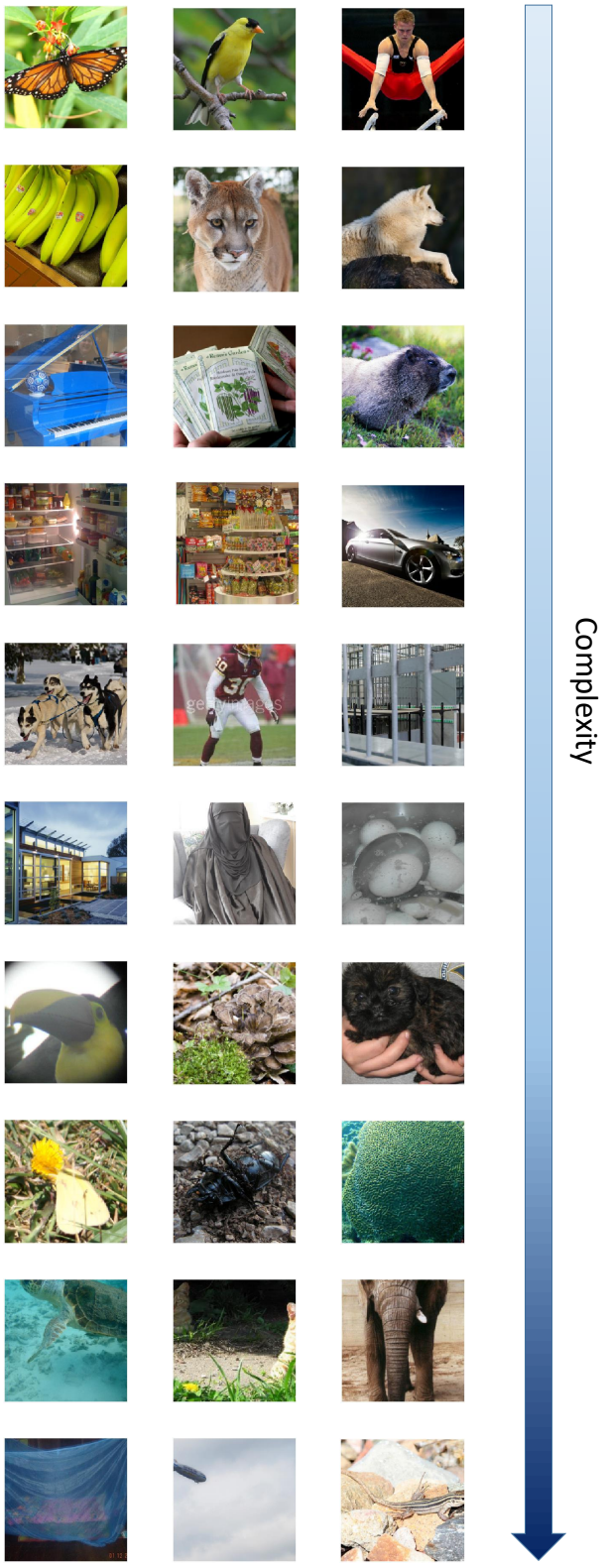


Figure S3. Images with different complexity on ImageNet. From top to bottom, the computational costs of sub-networks used to predict labels continue to increase.